Fouille de texte : une approche séquentielle pour découvrir des relations spatiales

Hugo Alatrista Salas*, Nicolas Béchet**

*UMR TETIS, 500 rue Jean-François Breton, F-34093 Montpellier
hugo.alatrista-salas@teledetection.fr
**IRISA, Université de Bretagne Sud, Rue André Lwoff, BP 573, 56017 VANNES Cedex
nicolas.bechet@univ-ubs.fr

Résumé. Dans cet article, nous présentons les premières étapes d'un projet de fouille de données textuelles. Plus précisément, nous appliquons un algorithme d'extraction de motifs séquentiels sous contraintes multiples afin d'identifier des relations entre entités spatiales. Les premiers résultats obtenus montrent l'intérêt de l'utilisation de cette approche et ses limites. Dans cet article, nous détaillons les premières bases de travaux plus ambitieux dont l'objectif est d'apporter des informations cruciales permettant de compléter l'analyse des images satellitaires.

1 Introduction

Les données satellites recèlent une multitude d'informations et leur analyse demande un investissement humain conséquent. La mise à disposition d'images avec une forte répétitivité temporelle pose le problème d'une gestion plus efficace, même si de nombreuses méthodes de classification automatique permettent d'aider l'expert dans cette tâche. De plus, les techniques actuelles de télédétection ne permettent pas une analyse efficace et rapide des images satellites dès qu'elles ont une forte répétitivité temporelle. Par exemple, grâce à la télédétection, nous pouvons déterminer des zones de cultures, mais nous ne pouvons différencier une culture de maïs ou de vignes pour un lieu donné. Il devient donc urgent de proposer des méthodes complémentaires à celles issues de la télédétection afin d'enrichir les connaissances associées aux images satellites. Dans ce contexte, l'utilisation de méthodes complémentaires à celles de télédétection telle que la fouille de texte semble pertinente.

Dans cet article, nous nous intéressons à la mise en place de méthodes d'extraction d'entités spatiales et de relations sémantiques entre ces dernières à partir de données textuelles. Actuellement, la plupart de ces méthodes s'appuient sur la mise en place de règles linguistiques (patrons lexico-syntaxiques). Seulement, étant donné le gros volume de données, de telles méthodes possèdent des limites et ne sont pas exhaustives. Dans le but de traiter une quantité importante de données, il semble essentiel de mettre en place des méthodes originales qui combinent des méthodes à base de motifs séquentiels et des approches de fouille de textes. Ce type de combinaison est proposé dans cet article.

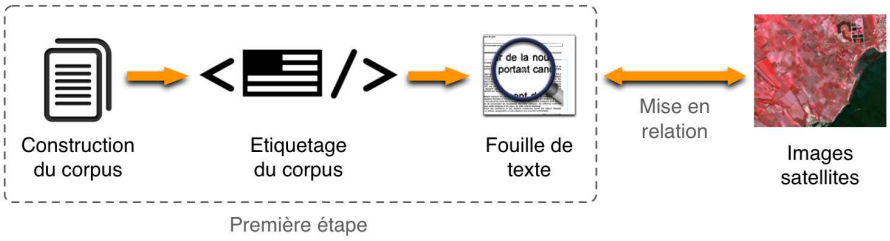

FIG. 1 – *Schema du projet de fouille de données textuelles*

2 Le corpus

Avec ses 7500 hectares, l'étang de Thau est, après la Méditerranée, la plus profonde et la plus grande étendue d'eau du Languedoc-Roussillon. Ce bassin s'étire le long de la côte lagune sur 19 km et 5 de large et plus de 30% du territoire de Thau est couvert par l'eau. Cet étang est un lieu dont le développement des activités économiques tels que la production d'huîtres et de moules ont augmenté énormément ces derniers années. Chaque année 13 000 tonnes d'huîtres, soit 10% de la production nationale sont extraits, ainsi que 3 000 tonnes de moules [1].

Dans un premier temps, nous nous intéresserons à des données textuelles issues du territoire côtier de Thau, déjà largement exploré par la communauté scientifique locale. Les données disponibles permettent de balayer les dispositifs numériques de cette région (journaux locaux, blogs citoyens ou d'associations, sites web des institutions locales, rapports, etc.) pour de premières extractions.

Le corpus à notre disposition est composé de 14 documents extraits du journal *Midi Libre* [2]. Il comporte des opinions, rapports et articles sur des différents rubriques tels que économiques, sociales, politiques, etc. de la région de *Sète*. Ces documents ont été nettoyés et postérieurement étiquetés en utilisant l'outil LinguaStrea [3] qui permet de traiter les textes et identifier les données géo-spatiales des *Entités Spatiales* (ES). La Figure 2 illustre un exemple d'étiquetage du corpus. Les balises *XML* illustrées dans la Figure 2 sont décrites ci-dessous :

- Sentence : l'identifiant de la séquence (une phrase) ;
- Token : peut être une proposition (PRO nous), un verbe (V avons), etc. ;
- Es : Les Entités Spatiales (ES) exprimées dans un texte sont constituées d'au moins une Entité Nommée (EN) et d'un nombre variable d'indicateurs spatiaux, précisant leur localisation. Généralement, une ES se réfère à une EN de lieu dont la position est connue (par exemple, Selles-sur-Cher, au nord de Jacou, dans les zones montagneuses) ;
- NomCommun : un nom commun, composé des différents Tokens, par exemple, (NOM l'ensemble)(PREDET du)(N territoire) ;
- Indirection : représente une position, par exemple, (PREP entre) ou (NOM est) ;
- VPE : comporte une expression incluant un verbe du type "perception", "position", etc. et un *NomCommun*, par exemple, (V vu)(CONJC que)(NOM l'espace)(A urbain).

1. Syndicat Mixte du Bassin de Thau - 328, Quai des Moulins - 34200 SETE.
2. http ://www.midilibre.fr/
3. http ://www.linguastream.org/

```
– <sentence id="n1" xml:id="d1p1s1">
    <token tag="CONJC">Comme</token>
    <token tag="PRO">on</token>
    <token tag="PRO">vous</token>
    <token tag="PRO">l'</token>
    <token tag="V">avait</token>
    <token tag="V">dit</token>
    ,
    <token tag="DET">un</token>
  – <nomCommun>
      <token tag="A">seul</token>
      <token tag="N">point</token>
    </nomCommun>
```

FIG. 2 – *Étiquetage du corpus : exemple*

Concernant les entités spatiales ES, cinq types d'indications spatiales existent : l'orientation (au sud de), la distance (à 1 heure de marche de, à 20 km de), l'adjacence (près de, loin de, la périphérie de), l'inclusion (le quartier de, la frontière entre, le sommet de) et la figure géométrique qui définit l'union ou l'intersection liant au moins deux ES (entre A et B, entre le triangle A, B, C, à l'intersection de A et B, la frontière A et B, etc.).

Dans la section suivante, nous décrirons la méthode d'extraction de motifs séquentiels utilisée dans nos expérimentations.

3 Extraction de motifs séquentiels : panorama et définitions

Le problème de la recherche de motifs séquentiels a été introduit par R. Agrawal dans [2] et appliqué avec succès dans de nombreux domaines comme la biologie [3, 4], la fouille d'usage du Web [5, 6], la détection d'anomalie [7], la fouille de flux de données [8], l'extraction de motifs spatio-temporels permettant l'étude des épidémies [9] ou la description des comportements au sein d'un groupe [10]. Des approches plus récentes [11] utilisent les motifs séquentiels pour décrire les évolutions temporelles des pixels au sein des séries d'images satellites.

Dans cette section, nous introduisons les définitions relatives à l'algorithme de fouille de motifs séquentiels sous contraintes multiples selon [1].

Dans l'extraction de motifs séquentiels, un *itemset* est un ensemble non vide des littérales appelés *items* noté par $I = (i_1, i_2, \ldots, i_n)$. Par exemple, *(NOM ville)* est un itemset composé des deux items : *NOM* et *ville*.

Une *séquence* S est une liste ordonnée, non vide, d'itemsets notée $< I_1 I_2 \ldots I_m >$ où I_j est un itemset. Par exemple, *S=<(NOM l'urbanisation)(ADJ ces)(ADJ dernières) (NOM années)>* est une séquence d'itemsets. Une *n-séquence* est une séquence composée de n itemsets. Par exemple, *S* est une 4-séquence.

Une séquence $S =< I_1 I_2 \ldots I_p >$ est une sous-séquence d'une autre séquence $S' =< I'_1 I'_2 \ldots I'_m >$, représentée par $S \preceq S'$, s'il existe des entiers $k_1 < k_2 < \ldots < k_j < \ldots < k_p$ tels que $I_1 \subseteq I'_{k_1}, I_2 \subseteq I'_{k_2}, \ldots, I_p \subseteq I'_{k_p}$. Par exemple, la séquence S' = *<(ces)(NOM années)>* est une sous-séquence de S car *(ces)* $\subseteq$ *(ADJ ces)* et *(NOM années)* $\subseteq$ *(NOM années)*. Toutefois, *<(ADJ)(NOM l'urbanisation)>* n'est pas une sous-séquence de S.

Une base de séquences *sBD* est un ensemble de tuples $(seqID, S)$ où *seqID* est l'identifiant de la séquence et S est une séquence. Par exemple, le Tableau 1 représente une base de séquences contenant trois séquences. Une tuple *(seqID, S)* contienne une séquence S_1 si et seulement si $S_1 \preceq S$. Le support d'une séquence S_1 dans la base de séquences sBD, noté par $supp(S_1, sBD)$ est le nombre de tuples dans la base de séquences qui contiennent la séquence S_1. Par exemple, dans le Tableau 1, Le support des sous-séquences *(PRO)(V)*, *(V correspond)(PREP à)* et *(PREP de)(NOM)* est 2 car ils apparaissent dans deux séquences différentes de la base de séquences.

Soit $minSupp$ le support minimal fixé par l'utilisateur, une séquence qui vérifie le support minimal (i.e. dont le support est supérieur à $minSupp$) est une *séquence fréquente*.

seqID	**Séquence**
S1	(PRO Nous)(V avons)(V vu)(CONJC que)(NOM l'espace)(ADJ urbain) (V correspond)(PREP à)(DET la)(NOM ville)
S2	(PRO Il)(V constituait)(NOM 11%)(PREP de)(NOM l'ensemble)(ART du) (NOM territoire)
S3	(NOM L'urbanisation)(V correspond)(PREP à)(DET une)(NOM expansion) (PREP de)(NOM 528)(NOM m2)

TAB. 1 – *Exemple d'une base de séquences*

Le problème de la recherche de motifs séquentiels dans une base de données consiste à trouver les séquences maximales dont le support est supérieur au support minimal spécifié. Chacune de ces séquences fréquentes est souvent appelée *motif séquentiel*. Le nombre de séquences fréquentes extraites peut s'avérer très important. La representation des séquences condensées tels que les *séquences fréquentes fermées* (closed patterns) [12] peut être donc utilisée pour éviter la redondance sans perte d'information. Une séquence fréquente S est dit fermée s'il n'existe pas une autre séquence fréquente S' tel que $S \preceq S'$ et $supp(S, sBD) = supp(S', BD)$. Par exemple, pour un support minimal $minSupp = 2$ la séquence <*(V correspond)*> du Tableau 1 n'est pas fermée, tandis que la séquence <*(V correspond)(PREP à)(DET)*> est fermée.

Le paradigme d'extraction de motifs séquentiels sous contraintes [13] apporte des techniques utiles pour inclure des contraintes choisies pour l'utilisateur (autres que le support minimal) afin de se concentrer sur les motifs les plus intéressants pour lui. Par exemple, il est possible de définir des contraintes pratiques telles que la contrainte de *l'écart*. L'écart (gap) dans une séquence est la nombre d'itemsets qui peuvent être ignorés compris entre deux itemsets d'une séquence fréquente S. $g(M, N)$ représente l'écart dont la taille est comprise dans l'intervalle $[M, N]$ où M et N sont des nombres entiers. L'intervalle $[M, N]$ est appelé *contrainte d'écart*. Une séquence fréquente satisfaisant la contrainte d'écart $[M, N]$ est dénotée par $P_{[M,N]}$. A savoir, il existe un écart $g(M, N)$ entre tous les itemsets voisins de $P_{[M,N]}$. Par exemple, dans le Tableau 1, $P_{[0,2]} = (V)(PREP)$ et $P_{[1,2]} = (V)(PREP)$ sont deux séquences satisfaisant la contrainte d'écart. Cependant $P_{[0,2]}$ est trouvé dans les séquences 1 et 3 tandis que la séquence $P_{[1,2]}$ est trouvé dans la séquence 2.

4 Extraction de motifs séquentiels textuels

A partir de documents textuels, ce travail a pour objectif d'identifier des relations entre les informations spatiales et tout autre mot contenu dans les documents. Dans ce cadre, une information spatiale est définie comme un nom toponymique précédé ou non d'un indicateur spatial et/ou d'une ou plusieurs relation(s) spatiale(s). Nous indiquons qu'un syntagme est lié à une information spatiale si la fréquence de la relation entre ce syntagme et l'ensemble des éléments de type information spatiale est supérieure à un seuil minimal fixé manuellement.

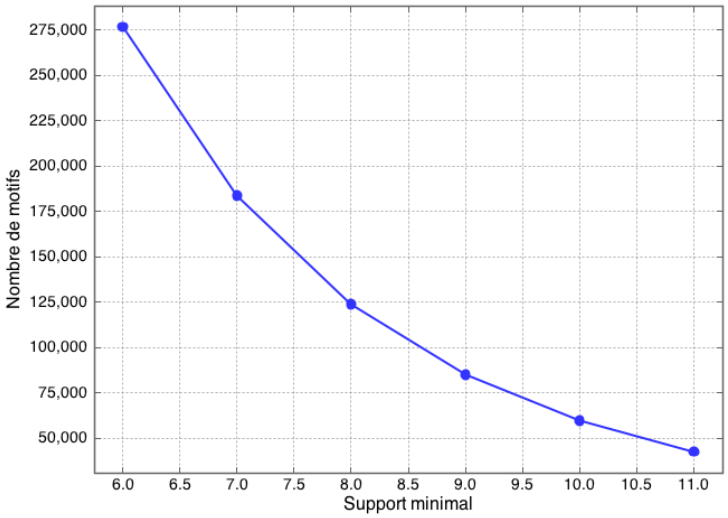

FIG. 3 – *Impact du support minimal sur le nombre de motifs extraits*

Les premières expérimentations ont été menées sur le corpus décrit dans la Section 2. Pour les expérimentations, nous avons appliqué un algorithme d'extraction de motifs séquentiels sous contraintes multiples (décrit dans la Section 3) où chaque phrase étiquetée du corpus est une séquence d'itemsets, par exemple, *(NOM L'espace)(ADJ urbain)...(PRE de)(TOPO)*. La base de séquences ainsi construite est composée de 241 séquences.

Pour les tests de cette première étape, nous avons expérimenté différents paramètres portant sur les contraintes :

- Le support minimal : plusieurs valeurs de seuil ont été utilisées. La Figure 3 représente l'impact du support minimal sur le nombre de motifs extraits. Les motifs fréquents illustrés dans le Tableau 2 correspondent aux experimentations avec un support minimal de 7 (un motif est inclus dans au moins 7 séquences de la base de séquences),
- Nous avons sélectionné les séquences (phrases) contenant au moins un toponyme caractérisé pour la balise TOPO,
- La contrainte d'écart (espace entre itemsets) a été fixée à 0, c'est-à-dire des itemsets consécutifs,

– La valeur de la longueur minimale (le nombre minimal d'itemsets comportant la séquence) a été fixée à 3,
– Les motifs séquentiels doivent obligatoirement contenir l'item *TOPO* (contrainte d'appartenance).

Les premiers résultats semblent cohérents, cependant, ils sont trop génériques même pour un support très faible. Il y a très peu d'itemsets dans les résultats et quand il y en a, il s'agit de termes très généraux comme "de" ou "et". Le résultat semble difficilement exploitable, d'autant que nous avons beaucoup de motifs (environ 183 000 motifs pour un seuil minimal de 7). Un outil de filtrage ou ranking de motifs (par exemple, les *top-k*) semble intéressant à mettre en place.

Les motifs qui finissent par des mots vides (conjonction, préposition, déterminant) ne sont pas intéressants. Pour résoudre ce problème, nous pouvons : a) faire un ranking en mettant en avant ceux avec une composition intéressante, par exemple, *(NOM)(ADV)(TOPO)* ; ou b) garder uniquement les préfixes des motifs, par exemple, *(NOM)(ADV)(TOPO)(DET)* devient *(NOM)(ADV)(TOPO)*.

Motif séquentiel	**Support**
<(TOPO)(DET)(NOM)(PREP)(DET)(TOPO)(NOM)>	7
<(TOPO)(DET)(PREP)(DET)(TOPO)(TOPO)>	7
<(TOPO)(DET)(PREP de)(N)(PREP)(TOPO)>	7
<(V)(NOM)(PREP)(NOM)(CONJC)(PREP)(DET)(TOPO)>	7
<(TOPO)(DET)(NOM)(DET la)(TOPO)>	8
<(DET)(NOM)(PREP)(NOM)(PREP)(TOPO)(CONJC)(DET)(TOPO)>	8
...	...

TAB. 2 – *Motifs séquentiels obtenus pour un support minimal de 7*

Finalement, pour limiter le nombre d'itemsets à analyser et évaluer, une solution est d'extraire uniquement les motifs qui se terminent par un toponyme *(TOPO)* car on trouve très souvent cette structure dans la langue française.

5 Conclusion

Dans cet article, nous avons présenté les premières étapes d'un projet de fouille de données textuelles. Nous avons plus particulièrement appliqué un algorithme d'extraction de motifs séquentiels pour extraire des relations spatiales. Ces relations spatiales seront postérieurement utilisés pour enrichir à faible "coût" humain les connaissances associées aux images satellites, plus précisément, nous allons déterminer les descripteurs linguistiques pertinents à partir des textes, puis de les mettre en correspondance avec les données images via l'identification des segments de textes du corpus spatialement pertinents pour une image donnée. Ce travail préliminaire suivra, dans un second temps, d'une généralisation de la méthode sur un large spectre couvrant les régions de Strasbourg, Pau, Caen et Montpellier offrant ainsi un corpus de données massives et hétérogènes.

Les perspectives de ce travail préliminaire sont nombreuses. Tout d'abord, nous devons valider la démarche proposée sur un jeu de données plus conséquent. Nous nous attacherons

également à proposer une mesure de validation objective permettant une selection des motifs obtenus.

Références

[1] Nicolas Béchet, Peggy Cellier, Thierry Charnois, and Bruno Crémilleux. Discovering linguistic patterns using sequence mining. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7181 of *Lecture Notes in Computer Science*, pages 154–165. Springer Berlin Heidelberg, 2012.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In Philip S. Yu and Arbee L. P. Chen, editors, *Proceedings of the Eleventh International Conference on Data Engineering, March 6-10, 1995, Taipei, Taiwan*, pages 3–14. IEEE Computer Society, 1995.

[3] Ke Wang, Yabo Xu, and Jeffrey Xu Yu. Scalable sequential pattern mining for biological sequences. In *CIKM '04 : Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management*, pages 178–187, New York, NY, USA, 2004. ACM.

[4] Paola Salle, Sandra Bringay, and Maguelonne Teisseire. Mining discriminant sequential patterns for aging brain. In Carlo Combi, Yuval Shahar, and Ameen Abu-Hanna, editors, *Artificial Intelligence in Medicine, 12th Conference on Artificial Intelligence in Medicine, AIME 2009, Verona, Italy, July 18-22, 2009. Proceedings*, Lecture Notes in Computer Science, pages 365–369, 2009.

[5] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, and Hua Zhu. Mining access patterns efficiently from web logs. In Takao Terano, Huan Liu, and Arbee L. P. Chen, editors, *Knowledge Discovery and Data Mining, Current Issues and New Applications, 4th Pacific-Asia Conference, PADKK 2000, Kyoto, Japan, April 18-20, 2000, Proceedings*, Lecture Notes in Computer Science, pages 396–407. Springer, 2000.

[6] Florent Masseglia, Pascal Poncelet, Maguelonne Teisseire, and Alice Marascu. Web usage mining : extracting unexpected periods from web logs. *Data Mining and Knowledge Discovery (DMKD)*, 16(1) :39–65, 2008.

[7] Julien Rabatel, Sandra Bringay, and Pascal Poncelet. Aide à la décision pour la maintenance ferroviaire préventive. In Sadok Ben Yahia and Jean-Marc Petit, editors, *Extraction et gestion des connaissances (EGC'2010), Actes, 26 au 29 janvier 2010, Hammamet, Tunisie*, Revue des Nouvelles Technologies de l'Information, pages 363–368. Cépaduès-Éditions, 2010.

[8] Alice Marascu and Florent Masseglia. Mining sequential patterns from data streams : a centroid approach. *Journal of Intelligent Information Systems*, 27(3) :291–307, 2006.

[9] Hugo Alatrista-Salas, Sandra Bringay, Frédéric Flouvat, Nazha Selmaoui-Folcher, and Maguelonne Teisseire. The pattern next door : Towards spatio-sequential pattern discovery. In *PaKDD'12, LNAI-7302*, volume 2, pages 154–168. Springer-Verlag, 2012.

[10] Dilhan Perera, Judy Kay, Irena Koprinska, Kalina Yacef, and Osmar R. Zaïane. Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6) :759–772, 2009.

[11] A. Julea., N. Meger, and Ph. Bolon. On mining pixel based evolution classes in satellite image time series. In *Proc. of the 5th Conf. on Image Information Mining : pursuing automation of geospatial intelligence for environment and security (ESA-EUSC 2008)*, page 6, 2008.

[12] Xifeng Yan, Jiawei Han, and Ramin Afshar. Clospan : Mining closed sequential patterns in large databases. In Daniel Barbará and Chandrika Kamath, editors, *Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, USA, May 1-3, 2003*. SIAM, 2003.

[13] Guozhu Dong and Jian Pei. *Sequence Data Mining*, volume 33 of *Advances in Database Systems*. 2007.